Bayesian View Synthesis and Image-Based Rendering Principles

Sergi Pujades, Frédéric Devernay, Bastian Goldluecke

Bayesian View Synthesis and Image-Based Rendering Principles

Sergi Pujades, Frédéric Devernay
Inria - PRIMA Team,
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France,
CNRS, LIG, F-38000 Grenoble, France.

Bastian Goldluecke
Heidelberg Collaboratory for Image Processing

Abstract

In this paper, we address the problem of synthesizing novel views from a set of input images. State of the art methods, such as the Unstructured Lumigraph [4], have been using heuristics to combine information from the original views, often using an explicit or implicit approximation of the scene geometry. While the proposed heuristics have been largely explored and proven to work effectively, a Bayesian formulation was recently introduced [28], formalizing some of the previously proposed heuristics, pointing out which physical phenomena could lie behind each. However, some important heuristics were still not taken into account and lack proper formalization.

We contribute a new physics-based generative model and the corresponding Maximum a Posteriori estimate, providing the desired unification between heuristics-based methods and a Bayesian formulation. The key point is to systematically consider the error induced by the uncertainty in the geometric proxy. We provide an extensive discussion, analyzing how the obtained equations explain the heuristics developed in previous methods. Furthermore, we show that our novel Bayesian model significantly improves the quality of novel views, in particular if the scene geometry estimate is inaccurate.

1. Introduction

We address the problem of novel view synthesis in the domain of Image-Based Rendering (IBR) [19], where the aim is to synthesize views from different viewpoints using a set of input views in arbitrary configuration. Most of the methods from the state of the art use heuristics to define energies or target functions to minimize, achieving excellent results. A major breakthrough in IBR was the inspiring work of Buehler *et al*. [4]. They define the "*desirable properties*" which any IBR algorithm should have. Those directives still prevail throughout the current state of the art.

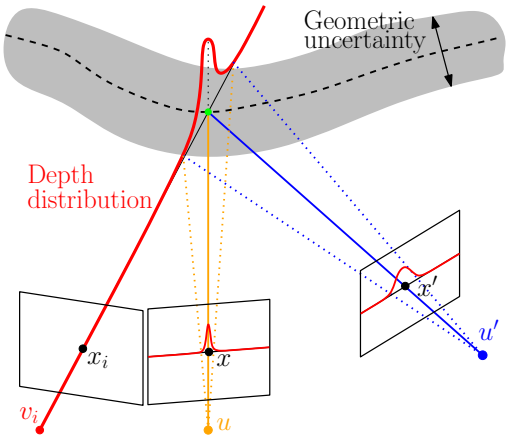

Figure 1. A depth distribution along an optical ray of camera v_i will propagate differently depending on the viewing angle of the rendered camera u or u'. The bigger the angle, the bigger the projected uncertainty will be.

Recently, however, the use of the Bayesian formalism has been introduced in IBR techniques, with the work proposed by Wanner and Goldluecke [28]. They provide the first Bayesian framework for novel view synthesis, describing the image formation process with a physics-based generative model and deriving its Maximum a Posteriori (MAP) estimate. Moreover, their variational method does not only address the problem of novel view synthesis. It directly addresses the synthesis of new super-resolved images, and provides a solid framework for other related problems, namely image denoising or image deblurring.

Interestingly, although [4] and [28] have addressed the same problem, their theoretical results do not converge into a unified framework. On the one hand, the guidelines dictated by Buehler *et al*. in [4] have proven to be very effective, but lack a formal reasoning supporting them. Moreover, it is unclear how the balance between some of the de-

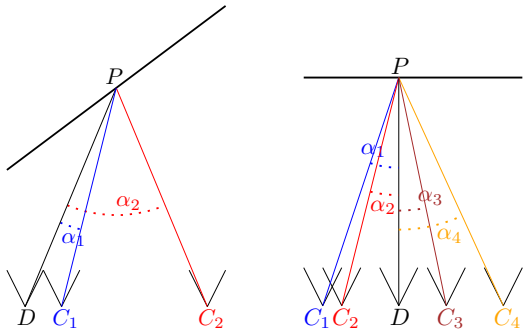

Figure 2. View D is generated from cameras C_i using [28]. Left: camera C_2 will be favored over camera C_1 because of the foreshortening effect. However, the angular distance of the viewing rays between D and C_1 is much smaller than D and C_2. Right: configuration with a flat scene. All cameras will have the same contribution, despite the different viewing angles.

sirable properties should be handled. An illustrative example is the tradeoff between *epipole consistency* and *resolution sensitivity*. The former notes that "*when a desired ray passes through the center of projection of a source camera it can be trivially reconstructed*", while the latter observes that "*in reality, image pixels are not really measures of a single ray, but instead an integral over a set of rays subtending a small solid angle. This angular extent should ideally be accounted for by the rendering algorithm.*" The *epipole consistency* is enforced with an angular deviation term, while the *resolution sensitivity* is driven by the Jacobian of the planar homography relating the views. Both heuristics seem reasonable, but which one should dominate? The choice of the weights between the properties is user-tuned, and in their experiments, parameters have to be adjusted differently depending on the scene.

On the other hand, the existing Bayesian model [28] is able to explain some of the heuristics, but still violates others which seem evident and have proven to work effectively. For example, we do find an analytic deduction of the influence of the foreshortening effects due to the scene geometry in the energy. The findings confirm the heuristic proposed by Buehler *et al*. in [4]: it is driven by the Jacobian of the transformation relating the views. However, when carefully analyzing the final equations in [28], an important desirable property proposed in [4] is still missing: the *minimal angular deviation* of the viewing rays is not enforced and even violated in some cases, as illustrated in Fig. 2.

The differences between state of the art generative models and the energies proposed by generally accepted heuristics is what has motivated the present work. Our goal is to retain the advantage of the intrinsically parameter-free energies arising from the Bayesian formalism, while pushing the image formation model boundaries of [28] and provide a new model which is capable of explaining most of the currently accepted intuitions of the state of the art in IBR.

Contributions. The key theoretical contribution of the proposed method is the systematic modeling of the error introduced in the lambertian image formation process via the inaccuracy in the estimates of the geometric proxy. We call this inaccuracy *depth uncertainty*, referring to the depth estimates from the input images. In addition to this error, we also consider the image sensor noise, commonly modeled as Gaussian. We extensively analyze the theoretical implications of the obtained energy, discussing the formal deduction of the state of the art heuristics from our model. This work provides the first Bayesian formulation explicitly deriving the heuristics of [4].

From a practical point of view, we numerically evaluate the performance of our method comparing it to the best existing method within the Bayesian framework. Experimental results show that we achieve state of the art results with regard to objective measures on public datasets. Moreover, we are also capable of addressing super-resolution, capitalizing on the general framework established in [28]. The new model is not without a price, since its optimization is less straightforward. However, existing methods allow us to overcome this difficulty. Source code is publicly available at `http://sf.net/projects/cocolib/`.

2. Related work

Since the early work on plenoptic modeling [14] proposed by McMillan and Bishop, many IBR techniques have been developed for several purposes, *e.g*. free-viewpoint rendering [24], image morphing [30] or image view interpolation [21] among others. The taxonomy done by Shum *et al*. [19] shows that most IBR methods use a geometric proxy, and they classify them in an "IBR Continuum" depending on how much geometry they use. On one end of this continuum we have methods which do not use any but rely on a large collection of input images, like light field rendering [13], and concentric mosaics [20]. On the opposite end, we have rendering techniques relying on explicit geometry, using accurate geometric models but few images, such as layered depth images [17, 6] and view-dependent texture mapping [9]. In between, we find methods using an implicit representation of the geometry, such as view interpolation techniques [7, 27] relying generally on optical flow, transfer methods [12] establishing correspondences along the viewing rays using epipolar geometry, and the Lumigraph [11], which uses an approximate explicit geometry and a relatively dense set of images.

When Buehler *et al*. introduced Unstructured Lumigraph Rendering [4], they established the seven "*desirable properties*" that all IBR methods should fulfill: *use of geometric proxies*, *unstructured input*, *epipole consistency*, *minimal angular deviation*, *continuity*, *resolution sensitivity*, *equivalent ray consistency*, and *real-time*. This work has been of major importance in the community, and most IBR methods

follow these guidelines.

Although Bayesian formalisms are a common way to deal with spatial super-resolution in the multi-view and light field setting [3, 10], they have only recently been introduced to IBR with the work by Wanner and Goldluecke [28]. While their work provides a physical explanation for the *resolution sensitivity* property, the *minimal angular deviation* can be violated in their final equations. Most interestingly, Vangorp *et al*. [26] empirically verify which properties in IBR methods are prone to create visual artifacts, and one of their main results identifies angular deviation as a key property to be taken into account to avoid visual artifacts.

Even if the performance achieved by state of the art 3D reconstruction methods in estimating geometric proxies is phenomenal, considering them as perfect seems too strong of an assumption: even the best ones have an uncertainty in their final estimates. Naturally, novel view synthesis is prone to producing visual artifacts in regions with a poor (implicit or explicit) reconstruction. One way to address this problem is to improve the acquisition setting, as done by Zitnick *et al*. [31]. They achieve a good enough reconstruction, leading to impressive novel view synthesis. However, their setting is heavily constrained.

In [22], Takahashi studies the theoretical impact of errors in the geometric proxy when rendering a new view from 2 images. We improve [22] by addressing more general camera configurations and providing an efficient method to find the solution, both explicitly left as future work. In [23] Takahashi and Naemura use the depth uncertainty information to leverage the regularizer term (prior). But this consideration does still not take into account the *minimal angular deviation*, because distinct contributions for each camera are not allowed. We solve this issue in this work.

3. Novel view synthesis generative model

Our goal is to synthesize a (possibly super-resolved) view $u : \Gamma \to \mathbb{R}$ from a novel viewpoint c using a set of images $v_i : \Omega_i \to \mathbb{R}$ captured from general positions c_i. We assume we have an estimate of a geometric proxy which is sufficient to establish correspondence between the views. More formally, the geometric proxy induces a backwards warp map $\tau_i : \Omega_i \to \Gamma$ from each input image to the novel view, as well as a binary occlusion mask $m_i : \Omega_i \to \{0, 1\}$, which takes the value one if and only if a point in Ω_i is visible in Γ. If we restrict τ_i to the set of visible points $V_i \subset \Omega_i$, it is injective and its left inverse $\beta_i : \tau_i(V_i) \to \Omega_i$ is well defined, see Fig. 3.

Ideal image formation model. In order to consider the loss of resolution from super-resolved novel view to input view, we model the subsampling process by applying a blur kernel b in the image formation process of v_i. It corresponds to the point spread function (PSF) of camera i. Each pixel of v_i stores the integrated intensities from a collection of rays

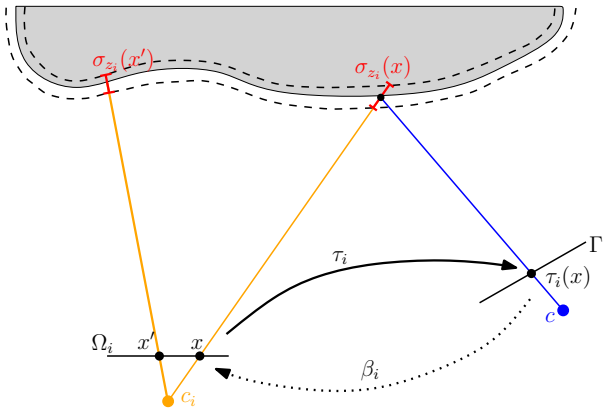

Figure 3. Transfer map τ_i from image plane Ω_i into target image plane Γ. The depth uncertainty σ_{z_i} may be different among pixels.

from the scene, and the novel view u will always be considered as having a higher resolution than the input views.

Let us discard the effects of visibility for a moment, supposing all points are visible. Also suppose we have a perfect backward warp map τ_i^* from Ω_i to Γ, and perfect input images v_i^*. Assuming the lambertian image formation model, the idealized exact relationship between novel view and input views is

$$v_i^* = b * (u \circ \tau_i^*), \tag{1}$$

being $\circ$ the function composition operator. However, the observed images v_i and geometry τ_i are not perfect, and we need to consider these factors in the image formation model.

Sensor error and image error. First, we consider Gaussian sensor noise on all cameras with variance σ_s^2. While the sensor noise variance σ_s^2 and the subsampling kernel b could be different among views, for the sake of simplicity of notation, we will assume them to be identical.

Second, we consider the error in the geometry estimate, which implies that the corresponding backwards warp τ_i is different from the ideal map τ_i^*. This induces an intensity error ϵ_{g_i} in the image formation process,

$$\epsilon_{g_i} = b * (u \circ \tau_i^*) - b * (u \circ \tau_i). \tag{2}$$

The uncertainty related to the intensity error ϵ_{g_i} is denoted by $\sigma_{g_i} : \Omega_i \to \mathbb{R}$. Note that both have intensity units.

Taking into account the above errors, the image formation model becomes:

$$v_i = b * (u \circ \tau_i) + \epsilon_{g_i} + \epsilon_s. \tag{3}$$

While we make the common assumption that ϵ_s follows a Gaussian distribution, the distribution of ϵ_{g_i} is yet unknown to us. What we know is that ϵ_{g_i} is strongly related to the geometric error. In the next section, we study the relationship between their distributions.

Dependency of image error on geometric error. The geometric proxy yields for each pixel x in Ω_i a depth measure z_i which is associated with an uncertainty σ_{z_i}, giving

us a distribution of depth along the viewing ray, as illustrated in Fig. 3. We now consider the error ϵ_{z_i} in the estimation of the geometric proxy, expressed in world units. The previous image error ϵ_{g_i} is dependent on the underlying geometric error. Note that the image error has intensity units and must not be confused with ϵ_{z_i} having geometric units. In contrast to the blur kernel and the sensor noise, we allow these errors to be different for each view and for each pixel in each view, as made explicit in the notation.

We assume that the error distribution for the depth estimates is normal, $\epsilon_{z_i} \sim \mathcal{N}(0, \sigma_{z_i}^2)$. The goal is now to derive how this distribution generates a color error distribution in the image formation process. Propagating a distribution with an arbitrary function is not straightforward, even if in our case, this depth error distribution is assumed to be Gaussian, and will only be propagated along the epipolar geometry between the views. Instead of computing the full color distribution along the viewing ray, we linearize and consider the first order Taylor expansion of v_i with respect to z_i. This implies that the resulting color distribution is also Gaussian, with mean $u \circ \tau_i$ and standard deviation

$$\sigma_{g_i} = \sigma_{z_i} \left| \frac{\partial v_i}{\partial z_i} \right| . \tag{4}$$

Using Eq. (1) and the chain rule, we find that

$$\sigma_{g_i} = \sigma_{z_i} \left| b * \frac{\partial (u \circ \tau_i)}{\partial z_i} \right| = \sigma_{z_i} \left| b * \left((\nabla u \circ \tau_i) \cdot \frac{\partial \tau_i}{\partial z_i} \right) \right| . \tag{5}$$

MAP estimate and energy. In the Bayesian formulation, the MAP estimate of the novel view can be found as the image u minimizing the energy

$$E(u) = E_{\text{data}}(u) + \lambda E_{\text{prior}}(u), \tag{6}$$

where the data term $E_{\text{data}}(u)$ is deduced from the generative model, and $E_{\text{prior}}(u)$ is a smoothing term which is detailed afterwards. $\lambda > 0$ is the only parameter of our method, and it controls the smoothness of the solution.

Let us consider the two error sources as independent, additive and Gaussian. Then their sum is also a normal distribution with zero mean and variance $\sigma_s^2 + \sigma_{g_i}^2$. The data term computed from the generative model of Eq. (3) is given by:

$$E_{\text{data}}(u) = \sum_{i=1}^{n} \frac{1}{2} \int_{\Omega_i} \omega_i(u)\, m_i (b * (u \circ \tau_i) - v_i)^2 \, dx, \tag{7}$$

$$\text{with } \omega_i(u) = \left(\sigma_s^2 + \sigma_{g_i}^2 \right)^{-1} . \tag{8}$$

This data term is similar to the one found in the previous model from [28], except for the factor $\omega_i(u)$, which can be seen as a weight that depends both on the depth uncertainty and on the latent image u being computed. If there were no depth uncertainty, this term would reduce to σ_s^2, which gives exactly the energy found in [28].

From Eq. (5), we can observe that the term $\sigma_{g_i}^2$ in $\omega_i(u)$ becomes smaller if the length of the vector $\partial \tau_i / \partial z_i$ decreases. The derivative $\partial \tau_i / \partial z_i$ denotes *how much the reprojection of a point x_i from the original view v_i onto the novel view u varies when its depth $z_i(x_i)$ changes*. This vector points towards the direction of the epipolar line on u issued from the point x_i of v_i, and its magnitude decreases with the angle between the optical ray issued from the original view v_i and the optical ray from the novel view u. As illustrated in Fig. 1, the term $\sigma_{g_i}^2$ thus accounts for the *minimal angular deviation* "*desirable property*" from [4], which was not accounted for in [28].

Let us analyze more precisely under which circumstances the weight $\omega_i(u)$ reaches its maximal value $1/\sigma_s^2$, which is the value found in the previous model. There are three situations in which this occurs. The first one is if $\partial \tau_i / \partial z_i = 0$, i.e. the depth of a point in v_i has no influence on its reprojection onto u. This can only happen if the two optical rays are identical, which corresponds to the *epipole consistency* property from [4]. The second one is if $\nabla u = 0$, *i.e.* the rendered image has no gradient or texture at the considered point: in this case, an error on the depth estimate has no effect on the rendered view. The last situation is if ∇u at the rendered point is orthogonal to the direction of the epipolar line from camera i passing through the rendered point: a small error on the depth estimate in camera i does not have an effect on the rendered view because the direction of influence of this error is tangent to a contour.

Choosing the prior. The prior is introduced in the Bayesian formulation to restrain the possible configurations of the target image. Usually, it is used to overcome the ill-posedness of the problem: in the analysis of super-resolution by Baker and Kanade [1], they show that the dimension of the null-space of the matrix system increases with an increase of the super-resolution factor. Furthermore, in novel view synthesis, some parts of the image may not be seen by any contributing view, thus a regularization prior allows to fill the gaps with plausible information. Thus, the choice of the prior will have significant influence on the final result.

Very interesting priors have been developed in order to overcome specific issues in super-resolution [18]. There are also techniques allowing to learn generic image priors from a collection of images [16]. However, in this work we focus on the generative model, and we use basic total variation as a regularizer,

$$E_{\text{prior}}(u) = \int_{\Gamma} |Du| , \tag{9}$$

which is convex and has been extensively studied in the context of image analysis problems [5]. The search for optimal priors will be a topic of future work.

Optimization. The energy from Eq. (6) is hard to opti-

mize because the weights $\omega_i(u)$ in Eq. (7) are a nonlinear function of the latent image u. Similarly to [8], we propose a re-weighted iterative method. We use an estimate $\tilde{u}$ of u, set at $\tilde{u} = \frac{1}{n}\sum v_i \circ \beta_i$ in the first iteration. We consider then $\omega_i(\tilde{u})$ constant during each iteration, making the simplified energy convex. Furthermore, with arguments similar to [28], we can show that the functional derivative of the simplified data term is

$$dE^i_{\text{data}}(u) = \omega_i(\tilde{u})\,|\text{det}D\beta_i|\,\left(m_i\bar{b} * (b*(u\circ\tau_i) - v_i)\right)\circ\beta_i, \tag{10}$$

where $\bar{b}(x) = b(-x)$ is the adjoint kernel. This functional derivative is Lipschitz-continuous, which allows to minimize the energy via the fast iterative shrinkage and thresholding algorithm (FISTA) [2]. With the solution of this simplified problem, we update $\tilde{u}$, thus obtaining new weights, and a new energy. We solve it again with FISTA, and iterate. Although the minimization problem to be solved within each iteration is convex, in general we cannot hope to find the global minimum of Eq. (6).

4. Relation to the principles of IBR

As we see in Eq. (10), the weighting factor for each view is composed of two terms. The term $|\text{det}D\beta_i|$ is the same as in [28] and corresponds to a measure of image deformation: it is the surface of a pixel from u projected to v_i. We can formulate the intuition behind it as *how much does the observed scene change when the viewpoint changes?*

The term $\omega_i(u)$ corresponds to the depth uncertainty, as was explained in the previous section. The intuition behind this is: *how much does the observed scene change if the measured depth changes?*

Let us now carefully establish the links of the proposed energy with the "*desirable properties*" of IBR stated in [4].

Use of geometric proxy & unstructured input. The geometric proxy is incorporated via the backward warp maps τ_i, and the input can be unstructured (*i.e.* a random set of views in generic position).

Epipole Consistency. As explained previously, the weighting factor $\omega_i(u)$ is maximal as soon as the optical rays from x_i and x are identical, so that if a camera has its epipole at x, then the contribution of this camera at x via the $\omega_i(u)$ term is higher. Epipole Consistency is thus satisfied.

Minimal angular deviation. This heuristic is provided by σ_{g_i} from Eq. (5): if all other dimensions are kept constant (resolution, distance to the scene, etc.), then the magnitude of the vector $\partial\tau_i/\partial z_i$ in Eq. (5) is exactly proportional to the sine of angle between the optical rays from both cameras to the same scene point.

Resolution sensitivity. This heuristic is followed by the term $|\text{det}D\beta_i|$, which measures the surface of a pixel from u projected to v_i. The larger the resolution of camera i, the bigger this surface, so that resolution sensitivity is properly handled.

Equivalent ray consistency. "*Through any empty region of space, the ray along a given line-of-sight should be reconstructed consistently, regardless of the viewpoint position (unless dictated by other goals ...)*" [4]. This is trivially satisfied by our framework, since the weights are varying continuously with the camera parameters (through the continuous variation of the backward warp maps τ_i). Moving the novel view camera along an optical ray (which is the situation used to describe this property in [4]) is just a special case.

Continuity. The *continuity* principle in IBR demands that the final rendered image is varying continuously with the camera parameters of the original views. This implies that there are no seams at visibility boundaries between cameras, which may happen near the borders of the intersection of the field of view of each camera with the scene, or at depth discontinuities seen from each camera. The typical heuristic to enforce this form of continuity is to lower the contribution of a camera near a visibility boundary or the boundary of its field-of-view [15, 4]. Our equations do not satisfy this property and the obtained weights do not fall to zero when approaching a visibility boundary. This could easily be enforced by smoothing the visibility maps m_i near the depth and visibility discontinuities, without changing the zero set of these functions. However, since we claim to have a completely physics-based Bayesian formulation, any operation on the visibility map should be sustained by a physical explanation, which we are still missing, and this is part of our future work.

Note that the prior term in the energy reduces the problems, most notably visual artifacts, which are due to not handling the continuity properly. However, a prior on the novel views cannot completely solve the continuity problem, which depends on the scene and camera geometry.

Real-Time. The final "*desirable property*" is for the method to be *real-time*. Our method is not yet real-time, mainly because of the computational complexity of the MAP estimate: 2 to 3 seconds are necessary to render a 768×768 image from 8 source images. However, both the resolution algorithms and the hardware architectures are evolving quickly, and much better performance can be expected in the next few years.

If super-resolution is not important, instead of solving the full MAP problem, it seems reasonable to use real-time regularization in the form of inpainting methods to obtain an acceptable result.

Balance between properties. One of the advantage of our method with respect to [4] is that the balance between the different properties is not handled by user-defined parameters, but implied from a formal deduction. Imagine a configuration with two cameras: one with low *minimal angular distance* but high *resolution sensitivity* change, and

another with high *minimal angular distance* but low *resolution sensitivity* change. Which one should contribute more to the final image? In [4], the *angular distance* is preferred to the *resolution sensitivity* by a ratio of $1/0.05 = 20$ (Hallway dataset). In our equations, these variations are completely physics-based. An angular deviation of $\Delta\alpha$ between views is penalized proportionally to $\frac{1}{\sin^2 \Delta\alpha}$, due to the change in $\sigma_{g_i}^2$. A foreshortening effect or resolution difference causing an image scale factor s is penalized proportionally to $\frac{1}{s^2}$, due to the change in $|\det D\beta_i|$. The balance between these factors is properly handled by taking into account the sensor noise σ_s^2.

An exception is the weight λ, used in the prior term. Note that this is common in all work on image analysis based on Bayesian principles: since there is currently no meaningful way to obtain a prior distribution on the space of images , one needs to work with regularization by objective priors. Of course one could also use existing methods [16] allowing to estimate this prior directly from the input images, thus obtaining a completely parameter-free model.

5. Experiments

Simplified camera configuration. Although we are addressing a generic case of novel view synthesis, in order to simplify the implementation of the optimization procedure, in the experiments we suppose that our cameras have a simplified configuration. Specifically, all viewpoints are in a common plane, which is parallel to all image planes, i.e. we are dealing with a 4D light field in the Lumigraph parametrization [11]. The novel view is also synthesized in the same image plane, which means that τ_i is simply given by a translation proportional to the normalized disparity d_i,

$$\tau_i(x) = x + d_i(x)(c - c_i). \tag{11}$$

Normalized disparity is expressed in pixels per world units, and is together with its associated uncertainty related to depth via:

$$d_i(x) = \frac{f_i}{z_i(x)} \text{ and } \sigma_{d_i}(x) = \sigma_{z_i}(x)\frac{f_i}{z_i(x)^2}, \tag{12}$$

where f_i is the camera focal length expressed in pixels.

Plugging (12) and (11) into (5), we derive the link between the geometric error and its associated image error as:

$$\sigma_{g_i} = \sigma_{d_i} \left|(b * ((\nabla u \circ \tau_i) \cdot (c - c_i)))\right|, \tag{13}$$

where σ_{d_i} models the disparity noise. Finally, the deformation term in Eq. (10) is:

$$|\det D\beta_i| = |\det D\tau_i|^{-1} = |1 + \nabla d_i \cdot (c - c_i)|^{-1}. \tag{14}$$

Datasets. To validate the theoretical contribution, we compare results on two light field datasets: The HCI Light Field Database [29], and the Stanford Light Field Archive [25]. These datasets provide a wide collection of challenging synthetic and real-world scenes.

In a first set of experiments, we render an existing view from the dataset at the same resolution, without using the respective view as an input to the algorithm. We consider two different qualities of geometric proxy: an approximate one from estimated disparity maps, and an extremely poor one represented by an infinite flat fronto-parallel plane in the estimated center of the scene. We adapt σ_{d_i} accordingly, *i.e.* when using the estimated disparity, we use a value corresponding to the expected accuracy of the reconstruction method: $\sigma_{d_i} = \frac{d_{\max} - d_{\min}}{\text{nbLayers}}$, where nbLayers is the number of disparities considered by the method. When a bare plane in the middle of the scene is used, we instead use $\sigma_{d_i} = \frac{d_{\max} - d_{\min}}{4}$. In all cases, $\sigma_s = 1/255$.

A second set of experiments is performed by rendering a 3×3 super-resolved image from a set of 5×5 input views. Although super-resolution is not the main purpose of the paper, we also provide a comparison with the state of the art. As super-resolution relies on sub-pixel disparity values, we only show the results obtained with the estimated disparity maps.

In Tab. 1, we show the numerical results obtained by our method, and compare it to the ones achieved with [28]. We measure the PSNR and DSSIM between the actual and generated images. Although our method visibly performs better, numerical values should be interpreted carefully. In Fig. 4, we show detailed closeups illustrating the benefits or our method. As high resolution images are not available for most of the datasets, PSNR and DSSIM values for the super-resolved images are computed by subsampling the input images, generating the novel super-resolved view and comparing it with the original one.

When rendering with precise geometry, both methods are roughly equivalent with respect to PSNR and DSSIM values (first and last two rows of Tab. 1). When the quality of the proxy degrades (third and fourth rows of Tab. 1), our method clearly outperforms previous work, taking advantage of the explicit modeling of depth uncertainty. As shown in the closeups of Fig. 4, our method better reconstructs color edges in all configurations. Full-resolution images are provided in the supplemental material.

Computation time when rendering at target resolution 768×768 with 8 input images is on the order of 2 to 3 seconds. Computation time for super-resolved view synthesis with a factor of 3×3 and 24 input images is around 2 to 3 minutes. All experiments used an nVidia GTX Titan GPU.

6. Discussion and conclusion

The main contribution of this paper is to establish the first formal link between the heuristics proposed in the recent decades for novel view synthesis, and the energy de-

	HCI light fields, raytraced				HCI light fields, gantry				Stanford light fields, gantry					
	still life		*buddha*		*maria*		*couple*		*truck*		*gum nuts*		*tarot*	
Estimated disparity														
Wanner *et al*. [28]	30.13	58	**42.84**	**17**	40.06	53	26.55	226	33.75	408	31.82	1439	28.71	60
Proposed	**30.45**	**55**	42.37	18	**40.10**	53	**28.50**	**178**	**33.78**	**407**	**31.93**	**1437**	**28.88**	**58**
Planar disparity														
Wanner *et al*. [28]	21.28	430	34.28	74	31.65	144	20.07	725	32.48	419	30.55	1403	22.64	278
Proposed	**22.24**	**380**	**37.51**	**44**	**34.38**	**99**	**22.88**	**457**	**33.79**	**386**	**31.30**	**1378**	**23.78**	**218**
Super-resolution														
Wanner *et al*. [28]	24.93	230	**34.50**	**122**	35.18	129	**25.54**	**287**	**33.11**	**378**	31.80	1475	**26.66**	**113**
Proposed	**25.12**	**228**	34.44	123	**35.20**	129	25.34	289	33.08	379	**31.89**	**1471**	26.54	117

Table 1. Numerical results for synthetic and real-world light fields from two different online archives. We compare our method to Wanner and Goldluecke [28] with respect to same-resolution view synthesis for estimated disparity and a flat plane proxy, as well as super-resolved view synthesis. For each light field, the first value is the PSNR (bigger is better), the second value is DSSIM in units of 10^{-4} (smaller is better). The better value is highlighted in bold. See text for a detailed description of the experiments.

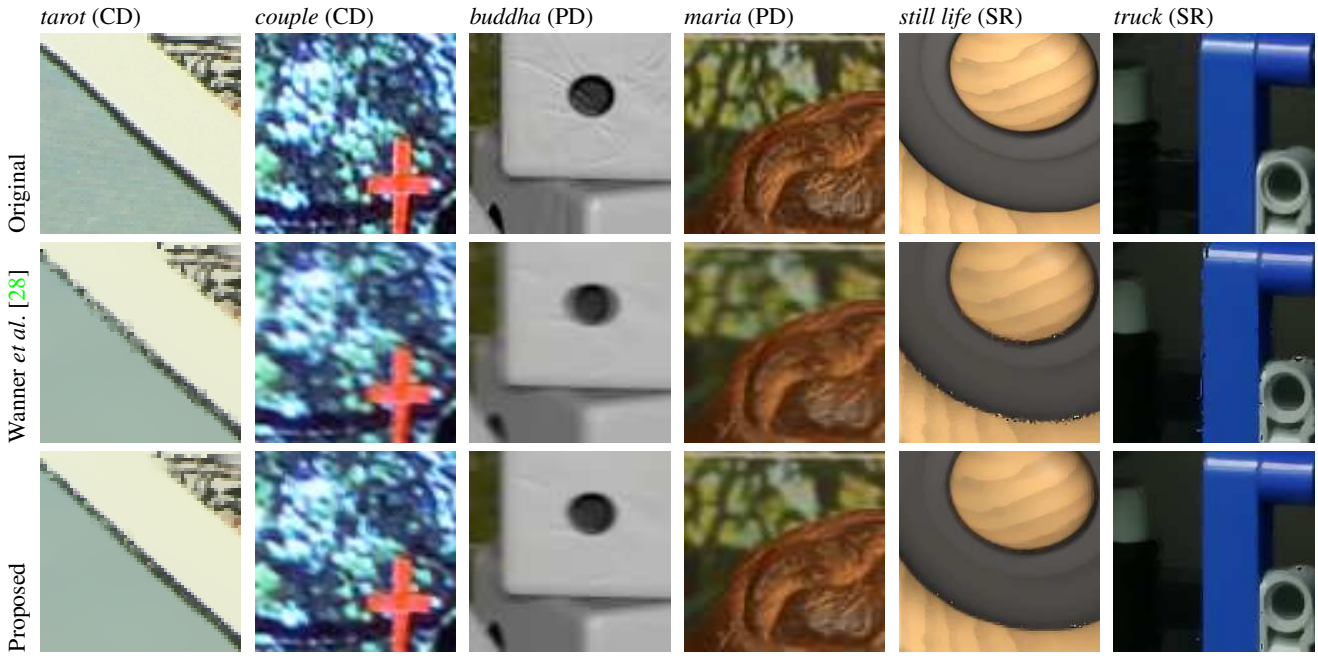

Figure 4. Visual comparison of novel views obtained for different light fields. From top to bottom, the rows present closeups of the ground truth images, the results obtained by [28], and our results. CD stands for computed disparity, PD for planar disparity and SR for super-resolution, see text for details. Full resolution images can be found in the additional material. The results obtained by the proposed method are visibly sharper, in particular along color edges.

ducted by a physics-based generative model.

This model can be used to solve the generic problem which consists in generating a novel view from a heterogeneous set of input images, and a geometric description of the scene (called a geometric proxy), which can be either explicit (*i.e*. the estimated geometry of the 3D scene) or implicit (*i.e*. a set of correspondence maps between original views and the novel view).

Part of our contribution is the analysis of how the proposed model fulfills almost all the guidelines established by Buehler *et al*. [4]. The proposed generative model provides a formal description of the intuitive heuristics behind these guidelines. The key element to this unification is to take into account the error in the estimated geometric proxy when rendering a new image. We have extensively discussed how our physics-based model explains the reasons why some important heuristics were picked up in the first place. The theoretical benefits of the model outperform state of the art by overcoming its limitations. Moreover, the experiments conducted on synthetic and real images show that our method improves state of the art performance in terms of rendered image quality.

Future work should better handle the visibility term in the model. In this work, visibility is computed from depth, but depth itself contains errors, which should propagate onto the visibility maps. This could be a key solution to incorporate the last missing *Continuity* heuristic into this physics-based Bayesian framework. Also extending the model to non-Lambertian scenes is crucial but quite hard. One would need to include general BRDF and lighting information to correctly model the transformation between input and novel views.

An important observation is that if the 3D reconstruction method or the 2D-2D image correspondence method provides not only depth estimates, but also the associated depth uncertainty, the image-based rendering method can benefit from this information to create better novel views. This should thus be a goal when developing new (implicit or explicit) reconstruction methods aimed at IBR.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *PAMI*, 24(9):1167–1183, 2002. 4

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIIMS*, 2(1):183–202, 2009. 5

[3] T. Bishop and P. Favaro. The Light Field Camera: Extended Depth of Field, Aliasing, and Superresolution. *PAMI*, 34(5):972–986, 2012. 3

[4] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured Lumigraph rendering. In *Proc. SIGGRAPH*, pages 425–432. ACM, 2001. 1, 2, 4, 5, 6, 7

[5] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1-2):89–97, 2004. 4

[6] C.-F. Chang, G. Bishop, and A. Lastra. LDI tree: A hierarchical representation for image-based rendering. In *Proc. SIGGRAPH*, pages 291–298. ACM, 1999. 2

[7] S. E. Chen and L. Williams. View interpolation for image synthesis. In *Proc. SIGGRAPH*, pages 279–288. ACM, 1993. 2

[8] T. S. Cho, C. L. Zitnick, N. Joshi, S. B. Kang, R. Szeliski, and W. T. Freeman. Image restoration by matching gradient distributions. *PAMI*, 34(4):683–694, 2012. 5

[9] P. Debevec, Y. Yu, and G. Borshukov. *Efficient view-dependent image-based rendering with projective texture-mapping*. Springer, 1998. 2

[10] B. Goldluecke and D. Cremers. Superresolution Texture Maps for Multiview Reconstruction. In *Proc. ICCV*, 2009. 3

[11] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The Lumigraph. In *Proc. SIGGRAPH*, pages 43–54. ACM, 1996. 2, 6

[12] S. Laveau and O. D. Faugeras. 3-d scene representation as a collection of images. In *Proc. CVPR*, volume 1, pages 689–691. IEEE, 1994. 2

[13] M. Levoy and P. Hanrahan. Light field rendering. In *Proc. SIGGRAPH*, pages 31–42. ACM, 1996. 2

[14] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *Proc. SIGGRAPH*, pages 39–46. ACM, 1995. 2

[15] R. Raskar and K.-L. Low. Blending multiple views. In *Proc. of Pacific Graphics*, pages 145–153. IEEE, 2002. 5

[16] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *Proc. CVPR*, volume 2, pages 860–867. IEEE, 2005. 4, 6

[17] J. Shade, S. Gortler, L.-W. He, and R. Szeliski. Layered depth images. In *Proc. SIGGRAPH*, pages 231–242. ACM, 1998. 2

[18] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. In *Proc. TOG*, volume 27, page 73. ACM, 2008. 4

[19] H.-Y. Shum, S.-C. Chan, and S. B. Kang. *Image-based rendering*. Springer, 2007. 1, 2

[20] H.-Y. Shum and L.-W. He. Rendering with concentric mosaics. In *Proc. SIGGRAPH*, pages 299–306. ACM, 1999. 2

[21] A. Smolic, K. Muller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems. In *Proc. ICIP*, pages 2448–2451. IEEE, 2008. 2

[22] K. Takahashi. Theory of optimal view interpolation with depth inaccuracy. In *Proc. ECCV*, pages 340–353. Springer, 2010. 3

[23] K. Takahashi and T. Naemura. Super-resolved free-viewpoint image synthesis using semi-global depth estimation and depth-reliability-based regularization. In *Advances in Image and Video Technology*, pages 22–35. Springer, 2012. 3

[24] M. Tanimoto. Overview of free viewpoint television. *Signal Processing: Image Communication*, 21(6):454–461, 2006. 2

[25] V. Vaish and A. Adams. The (New) Stanford Light Field Archive. `http://lightfield.stanford.edu`, 2008. 6

[26] P. Vangorp, G. Chaurasia, P.-Y. Laffont, R. W. Fleming, and G. Drettakis. Perception of visual artifacts in image-based rendering of façades. In *Computer Graphics Forum*, volume 30, pages 1241–1250. Wiley Online Library, 2011. 3

[27] S. Vedula, S. Baker, and T. Kanade. Image-based spatio-temporal modeling and view interpolation of dynamic events. *Proc. TOG*, 24(2):240–261, 2005. 2

[28] S. Wanner and B. Goldluecke. Spatial and angular variational super-resolution of 4D light fields. In *Proc. ECCV*, pages 608–621. Springer, 2012. 1, 2, 3, 4, 5, 6, 7

[29] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4D light fields. In *Proc. VMV*, 2013. 6

[30] G. Wolberg. Image morphing: a survey. *The visual computer*, 14(8):360–372, 1998. 2

[31] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *Proc. TOG*, volume 23, pages 600–608. ACM, 2004. 3